DOCUMENT RESUME

ED 187 727 TM 800 222

AUTHOR	David, Jane L.
TITLE	Local Uses of Title I Evaluations.
INSTITUTION	Stanford Research Inst., Menlo Park, Calif. Educational Policy Research Center.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C. Office of Planning, Budgeting, and Evaluation.
REPORT NO	EPRC 21
PUB DATE	Jul 78
CONTRACT	100-77-0095
NOTE	55p.
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Achievement Tests; Compensatory Education; Decision Making; Educational History; Elementary Secondary Education; Evaluation Needs; Informal Assessment; *Information Utilization; Program Attitudes; Program Effectiveness; *Program Evaluation; Program Improvement; *School Districts; Standardized Tests
IDENTIFIERS	*Elementary Secondary Education Act Title I; *Metaevaluation

ABSTRACT

A survey of administrators, teachers, and parents in 15 Elementary Secondary Education Act Title I school districts indicated that evaluations are used locally to meet federal and state requirements, to inform parents and staff, and to confirm positive attitudes toward a program. Standardized achievement tests, the backbone of Title I evaluations, are viewed as inadequate for judging or improving programs. Respondents felt tests were biased and preferred other measures: skill-specific tests, observation, self-concept or attitude measures. Evaluations were not used to improve programs for several reasons: program stability; funds; politics; unavailability of results in time for decision making; differing information needs of federal, local, and state agencies. Respondents disliked evaluation and ignored negative results if they believed in a program. Until these underlying attitudes change, evaluation results, even if technically sound, will not be used. Of the two current federal strategies for Title I evaluation, an independent national study and a local-to-state-to-federal reporting scheme, the latter is more likely to be used. The federal government should be committed to increasing communication between program staff and evaluation staff, and to promoting evaluation among local staff. (CP)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

ED182722

NATIONAL POLICY RESEARCH CENTER
Bob Interlandi

LOCAL USES OF TITLE I EVALUATIONS

RECEIVED
ERIC/TM

July 1978

Research Report EPRC 21

By: Jane L. David
Senior Policy Analyst

Prepared for:

Office of the Assistant Secretary for Planning and Evaluation
Department of Health, Education, and Welfare
Washington, D.C. 20201

Contract HEW-100-77-0095

SRI Project URU-6854

SRI International
EDUCATIONAL POLICY RESEARCH CENTER

3

MAR 3 1 1980

EXECUTIVE SUMMARY

Statement of the Problem

Title I of the Elementary and Secondary Education Act of 1965 (ESEA) was the first major social legislation to require program evaluation. The original requirement for Title I evaluations and its subsequent elaboration in the 1974 Amendments to the Act have resulted in a variety of interpretations of the purposes of the evaluations and several Federal strategies for their conduct. Since 1965, the Federal strategies for Title I evaluation adopted by the United States Office of Education (USOE) emphasize Federal information needs. By contrast, the legislative history of ESEA reflects a strong Congressional interest in the provision of evaluation information that is also useful for program improvement at the local level. The extent to which Title I evaluations have met Federal information needs has been studied, but there has been little attention paid to the impact of Federally mandated evaluations at the local level. This study was designed to investigate whether the same evaluation system can serve both local and Federal needs through an examination of local uses of Title I evaluation.

Objectives

This study was designed to answer two major questions: Do local staff use Title I evaluation results to identify strengths and weaknesses of their programs in order to improve them? Are the recent and proposed changes in the Title I evaluation system likely to alter local use of evaluation? Specifically, the study investigated how local Title I staff and parents use their Title I evaluation, what information they use in judging the effectiveness of their program, and how they make decisions about changing the program. The objective was to produce a report to document the history of Federal strategies in Title I evaluation, the uses of Title I evaluations by local staff and parents, the other types

of information used by local staff and parents in judging and in improving their program, and the implications of these findings for the current Federal Title I evaluation strategy.

Methodology

The primary sample consists of 15 Title I districts in six states. The districts were selected among those reputed to have an above-average emphasis on or concern with evaluation. The identification of such districts was based on recommendations of USOE staff, Technical Assistance Center directors, and state Title I directors. Although the sample is not nationally representative, choosing districts especially concerned with evaluation ensures that the findings are based on situations with the greatest potential for use of evaluations. In addition, the sample was augmented by field notes from another 15 districts collected in a concurrent USOE-funded study that involved interviews concerning evaluation in Title I districts.

The data collection consisted of face-to-face interviews with Title I administrators, principals of Title I schools, Title I teaching staff, and parents of Title I students. Copies of evaluation reports and other related documents were also obtained. A district visit was made by one or two interviewers for one to two days. The interviews were structured, to the extent that the same topics were pursued in each interview, but the emphasis on each topic and the specific questions were tailored to each situation and respondent.

The analysis consisted of drawing a tentative set of generalizations from several readings of the field notes. For each generalization, the notes were gone through carefully, extracting evidence in support of and opposed to the generalizations. After refining the general statements to be reported, quotations illustrating each point were pulled from the notes. From these lists, examples were selected for inclusion in the report, thus ensuring that the quotations reported in the text are indeed representative of the responses.

Major Findings

Posttest or gain scores reported for each project on standardized achievement tests comprise the main part of the district Title I evaluation for all the districts visited. Therefore, the findings often indicate uses of and attitudes towards standardized achievement tests rather than the evaluation report per se.

In general, the primary function the evaluations serve is to meet state and Federal reporting requirements; districts find that employing standardized achievement tests is the simplest way to meet these requirements. In addition, the evaluation is used to provide feedback to school staff and parents, often consisting simply of the provision of summaries of results to these audiences. Finally, respondents claimed that the evaluation report serves as an indicator of success, as a source of confirmation of existing beliefs about the program, and as a public relations document (however, these uses occur only when the results are positive).

From the responses to the direct question of how the evaluation results are used, it is clear that they do not primarily serve either as a means of judging the program or as a guide to program improvement. We pursued this issue in more depth by asking respondents how they judge programs and how program decisions are made.

From asking respondents how they would demonstrate that their programs are successful, and how they would make judgments about other programs, it is possible to deduce why evaluation plays such a limited role in these judgments. First, when local staff weigh standardized test results against other sources of information, such as skills-related tests and personal judgment or observation, the other sources of information almost always carry more weight. Second, a frequent explanation for ignoring evaluation results is that the scores are not meaningful because important background characteristics of schools (e.g., mobility) and children (e.g., socioeconomic status) have not been considered. Finally, in the eyes of staff and parents, the evaluation often excludes measurement of goals that they feel are as important as achievement, if not more. When asked what other types of information they would like for judging

programs, staff and parents typically cited curriculum-embedded and other skills tests rather than standardized tests, measures of noncognitive domains such as self-concept and attitude, and measures of program impact on parents, the community, and the staff itself.

As with judgments of programs, evaluation data are rarely mentioned in the context of decisions about program changes. Responses to questions about how program changes are decided suggest several reasons for this finding. One reason is that programs are quite stable; the changes that do occur tend to be marginal. Thus the universe in which to find connections between program change and evaluation is limited. A second reason is that the results of the evaluations are often not available in time for use in planning. A third reason is that factors other than evaluation, such as availability of funds and political concerns, play a major role in program decision-making. Finally, for the same reasons that evaluations are often ignored in judging program effectiveness (preference for other types of cognitive measures, belief in personal impressions, and concern with other outcomes), they are ignored in program planning. There are a few examples of changes in programs that were motivated in part by evaluation results, but these are the exception rather than the rule.

Stated reasons for not using evaluations tend to focus on the characteristics of the information they contain, and hence imply that if the type of information were changed, use of evaluation would increase. A careful consideration of respondents' statements as a whole, however, suggests otherwise. There are constraints on evaluation use imposed by the structure of Title I programs as well as unstated reasons for not using evaluation results, both of which must be understood in order to determine effective ways of increasing local use of evaluation.

Two of the constraints imposed by the structure of Title I programs have already been mentioned: the stability of programs and the timing of evaluations. Some other constraints were also observed, if not stated directly by respondents. First, in almost every district there is little connection between program staff and evaluation staff; this is particularly true in districts that use external evaluators. Consequently, there is

often little communication and understanding between those responsible for the administration and content of the program, on the one hand, and those responsible for the design and conduct of the evaluation on the other. Second, every Title I program contains multiple audiences with different information needs which are often overlooked in the design and reporting of evaluations. Finally, there is the general constraint imposed by the state of the art in educational treatments. Thus, deficits in knowledge about what constitutes a successful strategy in education limit the extent to which evaluations can be fully utilized. This constraint reflects not only the lack of proven alternatives, but also the frustration that the lack produces.

Beyond these contextual constraints, there are two attitudes of Title I staff that limit evaluation use. The first is that evaluation is usually perceived in a narrow and potentially threatening way. Evaluation is typically viewed as a set of procedures to provide one's superiors with information on which to judge the program, on the basis of criteria defined by those superiors. Hence, evaluation is more likely to be associated with accountability than to be regarded as a potential source of useful information. The second is that most Title I staff are deeply committed to the program; accordingly, they seek out evidence in support of their positive feelings toward the program and effectively ignore evidence that does not support these feelings.

These observations lead to the conclusion that changing the type or quality of information contained in Title I evaluations will not, by itself, significantly affect local use of these evaluations. To achieve an increase in local use, an evaluation system must attack the factors underlying the lack of use, both the elements of the program that act as constraints in themselves and the individual beliefs and attitudes that produce a negative view of evaluation.

Recommendations

USOE is currently employing two evaluation strategies. The first is a massive, multiyear study conducted by an independent contractor and designed to provide a national picture of the impact of Title I on

achievement. The second is the implementation of evaluation models within the three-tiered (local to state to Federal) reporting system designed to improve the quality and comparability of locally collected data. The second effort includes the provision of technical assistance from centers established in each of the ten HEW regions for this purpose. It is generally agreed that independent national studies provide the best source of evidence for the national impact of Title I. Because the best place to consider ways of increasing local use would seem to be within the three-tiered reporting scheme, my recommendations refer to this scheme.

First, any strategy designed to increase local use of Title I evaluations must be grounded in a Federal commitment to this goal--a commitment that must be understood and shared by the states and communicated clearly to local districts.

Second, districts need assistance in increasing communication and cooperation between program staff and evaluation staff. The site visits suggest that the provision of feedback can be used as one way to facilitate understanding between program and evaluation staff. However, the information fed back must be designed to be clearly understood by staff and parents and must meet the different needs of different levels within a district.

Third, Title I staff and parents need assistance in developing an understanding of the constructive role that evaluation can play as well as certain types of evaluation skills. Local staff have received little, if any, training in incorporating evaluation information into planning and decision making. In particular, they need assistance in learning how to ask their own evaluation questions. If the primary purpose of evaluation remains that of answering questions imposed externally, evaluations will continue to be perceived as potentially more threatening than helpful.

Until local staff view evaluation in a positive light, effort devoted exclusively to the development of technically sound data will be wasted in the context of local use. The USOE evaluation models are designed to improve the quality of the data and will not, by themselves, lead to an increase in local use of evaluation. However, the current

technical assistance strategy, if redirected, could serve as a powerful force in changing how evaluation is perceived and thereby increase evaluation use locally. To accomplish this goal, technical assistance must be redesigned to communicate a new view of the role of evaluation and to develop skills such as generating one's own evaluation questions. As long as technical assistance is defined narrowly as a way of telling local staff "how to improve the quality of their data," it will not increase local use of evaluations.

ACKNOWLEDGMENTS

The quality of any study based on personal interviews is dependent upon the cooperation and candor of the respondents. In this study, my staff and I were exceedingly fortunate to meet with individuals who not only responded willingly and openly to questions, but went out of their way to arrange complicated schedules, to provide information and to graciously play host to strangers. I owe tremendous gratitude to the many people at the local, state and Federal level, whose contributions form the backbone of this study.

I am also indebted to Susan Peterson and Henry Acland who conducted the interviews along with me. Their sensitivity and insights provided superb field notes from which to work and contributed substantially to the analysis. Henry Acland is due thanks as well for sharing with me the time-consuming task of extracting and classifying the quotations and reacting to embryonic stages of the report.

Significant improvements from the draft to the final report were made possible by the thoughtful comments of many reviewers and especially by the literary skills of David Greene. Finally, Keith Baker of the Office of the Assistant Secretary for Planning and Evaluation (DHEW) served as a model project officer providing valuable advice throughout the project on its design, implementation and particularly on the form and substance of the final report.

I am deeply appreciative of the assistance from all these people but cannot hold any of them responsible for the final product. The interpretations and conclusions are mine and do not necessarily reflect the views of either SRI International or the Office of the Assistant Secretary for Planning and Evaluation.

CONTENTS

EXECUTIVE SUMMARY iii

ACKNOWLEDGMENTS xi

I INTRODUCTION AND BACKGROUND 1

History 2
Design of the Study 8
Sample 9
Interviews 10
Analysis 11
Organization of Report 11

II PRIMARY USES OF EVALUATION 13

Meeting Requirements 13
Feedback of Results to Staff and Parents 15
Gross Index of Program Effectiveness 16

III USE OF EVALUATION IN JUDGING PROGRAMS 21

Limits of Evaluation in Judging Programs 21
Data Not Considered Persuasive 21
Important Variables Omitted 23
Important Goals Not Measured 24
What Information Is Used or Desired in Judging Programs . 25
Cognitive Growth 25
Noncognitive Outcomes 26
Areas Not Related to the Child 27

IV USE OF EVALUATION FOR PROGRAM IMPROVEMENT 29

Limits of Evaluation for Program Improvement 29
Program Stability 29
Irrelevance of Evaluations 30
Inappropriateness of Evaluations 31
What Types of Information are Used in Making Decisions . . 32

V DISCUSSION 37

Interpretation of the Findings 37
The Context of Title I programs 38
Underlying Attitudes Toward Evaluation 40
Conclusion 41
Implications for Policy 42

REFERENCES 47

I INTRODUCTION AND BACKGROUND

Title I of the Elementary and Secondary Education Act of 1965 (ESEA) was the first major social legislation to mandate evaluation. The legislative mandate was vague, reflecting a compromise between Robert Kennedy, who favored local accountability, and educational interest groups, who feared it. This mandate has resulted in a 13-year history marked by confusion and disagreement over the purposes of evaluation and the program itself. Senator Kennedy's original notion was that evaluation would provide parents and communities with information that could be used to press for reform. Thus, the original motive behind the evaluation requirements was a concern with the use of evaluation at the local level.

Since that time, numerous interpretations of the use of Title I evaluations have been put forth including a determination of the impact of Title I nationally, identification of successful programs, and the provision of information to local staff for improving their programs. The history of the Title I evaluation strategies adopted by the Federal government reflects an almost exclusive concern with Federal information needs. At the same time, the legislative history surrounding ESEA continues to reflect the view that Title I evaluations should also provide information useful at the local level in improving programs.

Can the same evaluation system serve both local and Federal needs? The extent to which Title I evaluations have met Federal information needs has been studied (McLaughlin, 1975); but little attention has been paid to the impact of Federally mandated evaluations at the local level--particularly in terms of their utility in providing information that can guide program improvement. Therefore, this study was undertaken to look specifically at local uses of evaluation. Do local staff use evaluation results to identify strengths and weaknesses of their programs in order to improve them? Are the proposed changes in the Title I evaluation system likely to alter local utilization of results? This study was designed to address these basic questions.

To set the stage for the design and findings of the study, it is helpful to review briefly the history of Federal strategies adopted for conducting Title I evaluations and to interpret their intent and success in meeting Federal and local information needs.

History*

Since the original legislation in 1965, each local educational agency (LEA) annually prepares an evaluation report for its state educational agency (SEA). Each state in turn compiles the results of the LEA reports and produces an annual evaluation report for the United States Office of Education (USOE). The first two years of this three-tiered reporting system were a major disappointment insofar as they did not produce consistent or comparable data that could be aggregated to provide a national picture of the effectiveness of Title I. As a result, while USOE was urged to improve the system and required by Congress (in a 1967 amendment) to report to them annually on the effectiveness of the programs, a somewhat different approach was instigated by the Office of the Assistant Secretary for Planning and Evaluation (ASPE). This approach can be characterized by its reliance on locally collected data, its production function approach to program effects, and its goal which was to determine for national purposes the elements of successful programs. The study, TEMPO (named after the division of The General Electric Company that conducted the study), was an acknowledged failure. The failure was unexpected because neither Federal officials nor researchers had faced the complexity of measuring program characteristics and costs, and because the desired achievement data were impossible to obtain.

During the next two years (1967-68 and 1968-69), USOE launched a third approach--an annual mail survey designed to obtain information on program characteristics, participant characteristics, and the achievement of participants. Again, however, the effort to describe the national

---

*The descriptions of the Federal strategies through 1970 are based in large part on information given by McLaughlin (1975).

impact of Title I on achievement was thwarted by the absence of usable achievement data. Only 5% of the survey responses included achievement data that could be analyzed.

Although the Title I evaluation efforts are described above as three approaches, they represent essentially the same strategy. They were all designed to provide a national picture of the impact of Title I on pupil achievement and they all rested on secondary analyses of locally collected data. They also shared the same fate in that they were all considered failures in producing the desired information. The failures were attributed primarily to the inadequacies of locally collected data for purposes of national aggregation; their value at the local level was not a subject of serious investigation.

Until 1971, the only exception to the above generalizations was a series of studies begun in 1968 and continued through 1971--the "It Works" series. Since the previous studies had not produced evidence of large achievement gains attributable to Title I nationally, there was a political need at the Federal level to demonstrate the success of Title I in raising achievement. Therefore, USOE commissioned the American Institutes for Research to conduct a search for exemplary programs--programs with evaluations showing substantial gains in achievement. The result was the identification of approximately 30 such programs, although later studies found that many of these projects either no longer existed or failed to demonstrate effectiveness.

In 1971, USOE adopted a new strategy in addition to the three-tiered reporting system. They began to collect primary data directly, instead of relying on data collected by LEAs. The Compensatory Reading Study, conducted by Educational Testing Service, was designed primarily to describe practices in compensatory reading programs, to assess their effectiveness in terms of achievement and their costs. The results of the Compensatory Reading Study showed little relationship between program participation and achievement and, like the former studies, was designed to produce a national picture, not to meet local information needs. Begun in 1971, the study was not completed until 1976 by which time the 1974 amendments to ESEA had again changed the Federal evaluation strategy.

Prior to 1974, legislative language did not reflect an explicit intent to have an evaluation system that produced information useful to local staff in improving their programs--a fact which is consistent with the early preoccupation with Federal needs. By the 1974 reauthorization, however, evaluation had become a major concern in Congress. In fact, evaluation activities in general had mushroomed from 1968 to 1973 as seen by the increase from $1.2 to $20.1 million in USOE planning and evaluation funds (GAO, 1977). The legislative history for this period reflects the multiplicity of purposes for the Title I evaluation system:

> The present law requires local school districts to conduct annual evaluations of their Title I programs and to report the results of these evaluations to the State educational agencies. The States in turn must submit periodic reports (including the results of the local evaluations) to O.E. The purposes of these requirements are: to enable each local educational agency to assess the effect of its program and to identify weaknesses as well as strengths of the project, thus serving as a tool for program revision and improvement; to enable each State to determine the extent to which progress has been made in reaching State goals for meeting the needs of educationally deprived children, as well as to provide a tool for State planning, management and dissemination; and to enable the Commissioner of Education to conduct a similar analysis at the national level. (Emphasis added.) (USCAN, 1974, p. 4111)

This intent of Congress, combined with the lack of nationally compelling data during reauthorization when the program was under attack from the administration, resulted in Section 151 of the 1974 amendments.

Section 151 went far beyond the preceding legislative requirements for evaluation in its specificity regarding evaluation and the responsibilities of USOE in conducting them. It contains the following requirements.

> Sec. 151.(a) The Commissioner shall provide for independent evaluations which describe and measure the impact of programs and projects assisted under this title. Such evaluations ... shall include, whenever possible, opinions obtained from program or project participants about the strengths and weaknesses of such programs or projects.
>
> (b) The Commissioner shall develop and publish standards for evaluation of program or project effectiveness in achieving the objectives of this title.

(c) The Commissioner shall, where appropriate, consult with State agencies in order to provide for jointly sponsored objective evaluation studies of program and projects assisted under this title within a State.

(d) The Commissioner shall provide to State educational agencies, models for evaluations of all programs conducted under this title ... which shall include uniform procedures and criteria to be utilized by local educational agencies, as well as by the State agency in the evaluation of such programs.

(e) The Commissioner shall provide such technical and other assistance as may be necessary to State educational agencies to enable them to assist local educational agencies in the development and application of a systematic evaluation of programs in accordance with the models developed by the Commissioner.

(f) The models developed by the Commissioner shall specify objective criteria which shall be utilized in the evaluation of all programs and shall outline techniques (such as longitudinal studies of children involved in such programs) and methodology (such as the use of tests which yield comparable results) for producing data which are comparable on a statewide and nationwide basis.

(g) The Commissioner shall make a report to the respective committees of the Congress ...

(h) The Commissioner shall also develop a system for the gathering and dissemination of results of evaluations and for the identification of exemplary programs and projects ...

The final part authorized funds for carrying out these provisions not to exceed 0.5% of the program's appropriation.

The initial response of USOE to this mandate was twofold: first, a contract was awarded to RMC Research Corporation which resulted in the development of evaluation models and second, a massive multiyear study, the Sustaining Effects Study, was designed. The former, in fact, had been designed prior to the passage of the 1974 Amendments in anticipation of the forthcoming legislative mandate. Its language was vague in that it did not specify evaluation models but rather asked for a "review and analysis of past reports and development of a model reporting system and format." The scope of work was expanded, however, and in 1975 produced three evaluation models. As described by RMC:

> The three evaluation models are: Model A, the Norm-Referenced Model; Model B, the Control Group Model; and Model C, the Special Regression Model. Each model has variations that enable it to be used with either normed (Model A1, Model B1, Model C1) or non-normed tests (Model A2, Model B2, Model C2).
>
> The norm-referenced evaluation design generates a no-treatment expectation from the assumption that the treatment group will maintain its status relative to the national norm group from pretest to posttest without treatment. The control group model utilizes the posttest (or adjusted posttest) scores of a control group as the no-treatment expectation. The special regression design employs the mean posttest score predicted from a comparison group's regression line as the no-treatment expectation. (RMC, 1976, p. 7)

Meanwhile, the plans for the Sustaining Effects Study awarded to Systems Development Corporation came under attack from Congressional staff. The study was undertaken for two purposes, according to USOE: to report on the numbers of economically and/or educationally deprived elementary school students who do and do not receive compensatory services; and to report on the benefits students derive from such services during more than one school year. The initial design was for a seven-year study at an estimated cost of approximately $25 million. According to Congressional staff, there was concern that with the bulk of the money going to a single national study, there would be little left to upgrade the evaluation capabilities of state and local districts as intended by the legislation. After negotiations between USOE and Congressional staff, the Sustaining Effects Study was reduced in scope and OE created Technical Assistance Centers (TAC) in each region, which began operating in October of 1976. The TACs provide free consulting services to SEAs, and through them to LEAs, on all aspects of Title I evaluation but particularly on the implementation of the evaluation models.

In summary, the two primary studies currently under way are based on strategies essentially the same as those in operation over the past six years: an independent national evaluation and the three-tiered reporting system. The difference is that considerable effort has been devoted to improving the methodology of both strategies. This is reflected in the longitudinal nature of the Sustaining Effects Study and in the evaluation

models proposed as part of the three-tiered reporting system, combined with the technical assistance on their implementation.

It is clear that national studies are not intended to provide information for local use, but rather to provide a national picture for USOE and Congress. Similarly, the development of evaluation models to be used in the three-tiered reporting system was motivated primarily by Federal needs. As the original Request for Proposal (RFP) for this development stated:

> This statement of work is intended to improve the data quality in LEA reports to states and State Title I Evaluation Reports submitted to USOE. In combination, these efforts should significantly improve the national data base upon which Title I impact is evaluated annually. (RFP 74-39, 1974)

The RFP went on to note that, for the purpose of obtaining national impact data, use of the on-going data collection efforts by LEAs and SEAs is less preferred than the use of data collected in a national study expressly for that purpose.

As this brief history demonstrates, Title I evaluation began with the idea of locally collected data passing up through a three-tiered system to the national level. The failure of this system to provide nationally useful data led to the current system, designed to impose procedures on LEAs to ensure that their data are compatible with national needs. Nevertheless, it is evident in the current reauthorization proceedings that there is still a strong desire on the part of Congress for data that can also be used locally. The House version of the bill (yet to go to conference) explicitly refers to local use in its evaluation requirements for LEAs, to wit:

> A local education agency may receive funds under this title only if ... the evaluations address the purposes of the programs ... and ... the results of those evaluations will be utilized in planning for and improving projects and activities carried out under this title in subsequent years. (H.R. 15)

In conclusion, the three-tiered reporting system is the only Federal strategy under way that carries the potential for providing locally useful data. Given the emphasis in the development of the proposed system

on the national need for data that can be aggregated, it is reasonable to ask whether the new system will meet local needs better than its predecessors. Therefore, this study was designed expressly to investigate the extent to which the Title I evaluation system has been providing data that are used by local Title I staff in planning for and improving their programs; and to anticipate the impact of the new evaluation models and Technical Assistance Centers on the local utility of the data collected under this system.

Design of the Study

The design of this study was influenced by some recent research on the connections between evaluation findings and decision-making (for example, Weiss, 1977; Cohen and Garet, 1975; Frankel, 1976). This influence was primarily one of limiting expectations, which in turn influenced the approach to data collection and the selection of districts to be visited. Literature on evaluation has only recently included attempts to understand the role and use of evaluation results, particularly in the realm of program planning and decision-making. Partly in response to the absence of compelling evidence of evaluation utilization in decision-making, this research has tempered idealistic notions of clear connections between evaluation results and decisions. It has begun to suggest the bounds within which evaluation can reasonably be expected to provide usable information, and shows that the role of evaluation in decision-making can be important even if indirect and elusive (for example, in setting a climate of opinion or lending weight to common sense understandings about programs).

On the basis of these findings and my own work in local school districts, I embarked on this study with highly restrained optimism about being able to identify uses of Title I evaluation results. I did not expect to find very many examples of use of evaluations, either in judgments about programs or in decisions about changing programs. Therefore, I decided first that it was essential to interview Title I staff and parents in person, allowing enough flexibility to adapt questions to the respondents and their situation and to probe into each LEA's decision-

making process. I also decided consciously to select districts reputed to have an above-average emphasis on or concern with evaluation results.

Sample

To identify districts that emphasized evaluation, I asked knowledgable persons for recommendations, including directors of the Technical Assistance Centers, USOE staff, other researchers, and local staff with whom I was already acquainted. For most of the sample, the final selection was based on recommendations of the state Title I directors in states suggested to me. After explaining the purposes of the study, the state directors suggested several districts that met my criterion, from which I chose those to be visited.

We visited 15 districts in 6 states: 3 in California, 3 in Washington, 3 in West Virginia, 3 in Iowa, 2 in Nebraska, and 1 in New Mexico. Of the 15 districts, 10 were small to medium-sized cities ranging in population from approximately 75,000 to 500,000, with a median of approximately 200,000. The remaining 5 districts were rural to suburban with populations ranging from approximately 10,000 to 200,000.

I was able to augment the sample with an additional 15 districts through the cooperation of the Huron Institute and USOE. The Huron Institute was concurrently collecting similar information from local districts in their USOE-funded study on the feasibility of developing evaluation models for Title I early childhood programs. Their sharing of findings in effect doubled the sample and expanded the range by representing an additional 6 states and including 4 cities with populations between 500,000 and 1 million.

Both the selection procedures for the sample of districts and the size of the sample clearly preclude statistically valid generalizations to the nation as a whole. Choosing districts especially concerned with evaluation, however, ensures that the findings are based on situations with the greatest potential for use of evaluation. Therefore, these districts should represent the high end of the continuum of evaluation use in program judgments and decisions. Similarly, conclusions concerning

factors inhibiting evaluation use will apply even more to districts with less emphasis on evaluation not represented in the sample.

The sample also included four state Title I offices. For the purposes of this report, the findings from the state-level interviews serve as a background for the interpretation of findings from the districts. Because of time and budget constraints, however, the information gleaned from the states is not reported here. I hope to expand upon the data base and report on state-level findings in the future.

Interviews

A district visit was made by one or two interviewers for one to two days. In each district, we interviewed the Title I director, other project administrators, the Title I evaluator, principals of Title I schools, Title I teaching staff and parents of Title I students. In some districts, non-Title I administrators, such as the superintendent, were also interviewed. Title I director and Title I evaluator are my terms for the persons responsible for the administration and evaluation of the program. Their actual titles varied from district to district as did the titles of other Title I administrators.

In each district, the interviews were set up by either the Title I director or evaluator and were done either individually or in small groups, depending upon scheduling convenience. Generally, the interviews lasted from one-half hour to one hour and occurred either in the central office or at the school sites. The interviews were structured, to the extent that the same topics were pursued in each interview, but the emphasis on each topic and the specific questions were tailored to each situation and respondent. The categories of topics included:

Characteristics of the Title I program
How program decisions are made
Characteristics of the local Title I evaluation
Knowledge of the local Title I evaluation
Uses of evaluation results in judging programs and in program planning
Knowledge of and reactions to the evaluation models and TACs.

We chose not to tape record the interviews in order to maximize candor on the part of the respondents. However, we did take extensive notes, including as many verbatim quotations as possible.

In addition, in each district we obtained copies of their evaluation reports and other related documents.

Analysis

The task of synthesizing approximately 1,000 pages of typed field notes is awesome. The approach consisted essentially of reading the notes several times and tentatively drawing a set of generalizations from them. For each generalization, the notes were gone through carefully, extracting evidence in support of and opposed to the generalization. After this stage of refining the general statements to be reported, statements illustrating each point were extracted from the notes. From these statements, examples were selected for inclusion in the report. Thus each quotation reported in the text represents a much larger set of quotations illustrating the same point. This procedure was followed to ensure that the quotations were indeed representative of the districts. Since the interviews were not taped, the quotations are not demonstrably verbatim. They do, however, reflect the words of the respondents as closely as possible and capture the flavor of the response. For this reason, the vast majority of the quotations are based on the 15 districts we visited personally; and those from the Huron Institute sample are used sparingly.

Organization of Report

The body of the report is composed of three sections; primary uses of evaluation, uses of evaluation in judging programs, and uses of evaluation for program decisions. These sections are intended to be primarily descriptive; however, the urge to interpret has not been completely controlled. The final section contains interpretations and conclusions drawn from the findings and their implications for the future of Title I evaluation.

II PRIMARY USES OF EVALUATION

I found that the main part of the district Title I evaluation report for all the LEAs visited consists of posttest or gain scores reported for each project on standardized achievement tests. A few evaluations included additional information, such as the results of questionnaires given to staff and parents soliciting their opinions of the project. On the whole, however, program evaluation is synonymous with standardized achievement test scores. Accordingly, the findings presented throughout often indicate uses of and attitudes toward standardized achievement tests rather than the evaluation report per se.

This section presents the responses to the general question: How is the Title I evaluation used in your district? The most frequent responses fall into three areas: to meet requirements, as feedback to school staff and parents, and as a rough index of the program's impact on achievement.

Meeting Requirements

There is little doubt that the primary function the evaluations serve is to meet the state and Federal reporting requirements of Title I. Districts employ standardized tests because they are the simplest way of meeting the Federal mandate as interpreted by their state. LEAs are totally accustomed to the fact that receiving Federal money has a number of strings attached to it, of which the evaluation requirement is merely one. For example,

> This district will accept all strings that go with the Federal money. Richer ones might not but we need the money.
> (Director)*

---

*Throughout the text the type of respondent is identified in parentheses. All directors, administrators, evaluators, teachers, and parents are part of Title I. Principals are all in Title I schools and non-Title I administrators are all superintendents.

Therefore, with the exception of staff concerns with the time devoted to testing and the reporting burden, evaluation is usually perceived as just one of the many hoops to go through in order to receive the funds.

> We go along with externally imposed regulations as long as they do not impose an overwhelming burden. When they are burdensome, we will exercise our own judgment about what is legitimate and not go down without a fight. (Evaluator)

> Evaluation is not a burden; it is an unnecessary but required evil. It does little harm but is of no particular use. (Teacher)

So long as the burden is not undue and some local autonomy is preserved in designing their program, most local staff responsible for conducting the evaluation are concerned primarily with meeting the legal requirements:

> Testing is an economical and straightforward way of complying with the regulations; we send the data in and then go about our business. We're not going to lose any sleep over whether or not the results show effectiveness. (Evaluator)

Providing data to meet evaluation requirements is an accepted fact of life. Title I staff also believe that the Federal government has a right to request the data because they are footing the bill. Moreover, many but not all Title I staff think that there is a real need for the data at higher levels (i.e., district, state, or Federal). One district director described the perceptions of his staff in the following way:

> Teachers feel that all this data collection goes on because the state needs it or more generally the government needs it and they are sympathetic with their need for knowing what happens with their money. But outside of this necessity, they see little purpose. (Director)

Similarly, in another district,

> There is a real need for the big picture at the state and national levels. (Director)

In the context of the new USOE evaluation models, the district evaluator said:

> I can see the Federal and state need to demonstrate bang for the buck but cannot see why they avoid educators in coming up with guidelines. (Evaluator)

Some other respondents, however, were less sanguine about the appropriateness of aggregating these data for national purposes. For example,

> I doubt that the three-tiered scheme will give the feds what they want. A national picture is not appropriate. You have to accommodate to too many differences and the accommodations wash out the differences. (Evaluator)

Feedback of Results to Staff and Parents

The second primary use of evaluation results is to provide feedback. Feedback in this context connotes simply communicating evaluation results to program staff and parents. Theoretically, this is the area that provides the greatest potential for use of evaluation in making judgments and decisions about programs leading to improvements. As one district administrator stated:

> If the test data are not useful to the principals, they aren't useful at all. (Administrator)

All districts provide some type of feedback, but the type of information fed back varies enormously. At a minimum, feedback consists of sending the evaluation report to the Parent Advisory Council (PAC) and the principals of Title I schools. This situation is the one least likely to lead to any utilization (or even understanding) of the information. Principals rarely look at the report under these conditions, and teachers often do not see it. Most districts, however, provide school by school results, and sometimes class level results, which are transmitted to the appropriate individuals.

Sometimes this information is quite comprehensive. For example, in one district each Title I school receives a 15 page mimeographed document containing graphs of the relationship between school level poverty indices and achievement (with the particular school's code circled), detailed test score results for the school (by subtest and skill area), with national percentiles, and a comparison with the previous year's data for that school. It also contains other descriptive data on school and community characteristics such as mobility, enrollment, and income. The introduction to the report reads:

> The purpose of this report is to share information about students in Title I schools in _____. It is intended that the report be seen not as an evaluation report but as a collection of information that will help administrators, teachers and parents plan even stronger programs for the children in these schools.
>
> Much of the information reported here was collected as part of the data base used to evaluate ____ Title I programs. In addition, the _____ Research and Testing division contributed data it has gathered through the state mandated testing program.

This district was extraordinary in the efforts put forth by the evaluation staff to make evaluation part of the program planning effort. They go to considerable effort to present the information for each school clearly and to explain the findings in person to teachers, parents and the principal. In this district, as well as others, it was stated that feedback that included personal explanations by evaluation staff was much more likely to be understood and make an impression on the school staff; and hence have the potential to be utilized. In another district, the director said:

> Principals won't make any use of evaluation results if you just send data--you need to go talk with them about it. (Director)

In sum, the provision of feedback, particularly when explained in person, provides what may be a necessary but not sufficient condition for utilization of the evaluation information.

Gross Index of Program Effectiveness

Although the primary local uses of evaluation are to meet requirements and to provide feedback, other uses are not precluded. The third major use falls under the category of gross or rough indications of program effectiveness. This category differs from the previous ones in that it occurs at the individual rather than the system level. The use of evaluation as a gross index of program accomplishments takes several forms, the most common of which is use of evaluation to confirm one's existing beliefs about a program. For example:

> The main purpose the test scores serve is to support your own views. (Teacher)

> I look at test scores mainly to confirm my own impression. If they differ, my impression counts. (Teacher)

A related use of evaluations under this category is that of giving a rough index of program success, but not as a guide to action. For example,

> Standardized achievement tests provide an indicator of where the children are. (Teacher)

> Tests can only be interpreted as a rough guide. (Principal)

Also related to this category is the use of evaluation as a public relations document.

> I want information to justify expansion of the program. I'm not interested in information showing students are behind national norms. (Superintendent)

Illustrating another form of public relations, in one district the evaluator explicitly pointed out the need to use the evaluation report as a way of educating the district administration and board to have realistic expectations about the effectiveness of their Title I program. Similarly, in another district the reading program director described the situation in which a school wanted to withdraw from participation in Title I:

> They claimed that Title I was associated with a decline in test scores. We were able to pull out the evaluation report and demonstrate that this was not true. (Administrator)

In another district, the superintendent stated:

> Day-to-day problems don't show up in the evaluation. Subjective feedback is often more useful in daily operation of the program. The other stuff is what you impress people with. (Superintendent)

These common uses of evaluation--as a source of confirmation of existing beliefs, as an indicator of success, and as a public relations document--share an important characteristic: they are triggered only by positive results. Thus, the evaluation report as an end in itself (apart from meeting requirements) is seen as useful only when the results are positive. When the results are negative, the evaluation is discounted

for any number of reasons. (Elaboration of these reasons is contained in the following section.) Thus, it is often the case that negative findings, rather than being taken as informative about the program, are viewed as an annoyance that must be explained away. As two examples,

> None of the subjective stuff is included in the evaluation except occasionally to explain low results. (Superintendent)

> One year the scores for second grade were low. We looked for the reason by talking with the teachers to see if the skills tested matched the curriculum and if the students' scores matched the teachers' judgment. From this, we concluded that the test was invalid. (Evaluator)

In another district, the district staff were all extremely upset over results that showed negative NCE growth.

> We are having the TAC reanalyze our data looking for floor effects. We know instructional growth is taking place. The negative results hurt us in several ways. First, Congress is always talking about the possibility of tying funds to gains and, second, we get a bad reputation. Poor results limit our ability to share information about the program and lead to low morale. (Director)

And in another district, one school had very low scores:

> We discovered that there had been an influx of Vietnamese students into the school. In another school with low scores we found that there were a number of students who were near EMR (educable mentally retarded). (Director)

However, not all evaluations with negative findings are ignored. There are a few instances in which they are taken as a gross indicator of a weakness, but this occurs primarily in the context of needs assessment. Although the same set of standardized scores, or at least the same type, are used for both needs assessment and evaluation, they are far more likely to be seen as useful and acted on when they are viewed as needs assessment data as opposed to evaluative data. This point is expanded upon more under the section on what information people claim they use.

Ironically, there is almost unanimous agreement that standardized tests (especially when combined with teacher judgment)* form a good basis for selecting students for the programs. Although this use does not relate specifically to evaluation, it is mentioned here because it is a widely approved "good" use of standardized tests in a world in which they are usually criticized severely.

In summary, the primary local uses of Title I evaluations are to meet legal requirements, to provide feedback, and to provide gross indicators of program effectiveness. Title I evaluations do not seem to serve, as primary purposes, either as a basis on which to judge the program or as a guide to program improvement. Since direct inquiries about uses of evaluation results did not reveal use in program planning and improvement, we pursued the issue in more depth by asking respondents how they judge the programs and how decisions about programs are made. The findings from these inquiries are reported in the next two sections.

---

*Two districts were exceptions. One felt strongly that teacher judgment should not be included and another based selection exclusively on teacher judgment.

III USE OF EVALUATION IN JUDGING PROGRAMS

This section discusses evaluation in the context of judging the effectiveness of programs. Everyone involved in a Title I program makes judgments about its effectiveness. These judgments can be an end in themselves or they can be the basis for deciding how to change the program in order to improve it. Evaluation is viewed somewhat differently from these two perspectives; therefore, I treat these two perspectives separately. Section IV presents the findings on evaluation in the context of program planning and redesign.

Before discussing the ways in which evaluation results do affect program judgments, it is useful to consider the reasons that limit their utility.

Limits of Evaluation in Judging Programs

From asking respondents how they would demonstrate that their programs were successful and how they would make judgments about other programs, it is possible to deduce why evaluation plays such a limited role in these judgments. There are three classes of reasons limiting the impact of evaluation on judgments of program success: the data they provide are not considered as persuasive as other sources of information; the analyses ignore important mediating variables; and the evaluations don't measure important goals.

Data Not Considered Persuasive

When local staff weigh standardized test results against other sources of information in judging the success of their program, the other sources of information almost always carry more weight. Conflicting information from standardized tests and sources such as criterion or other skills related tests and personal judgment (gleaned from

observation, intuition or some combination) are inevitably resolved in favor of the other sources. Some examples follow:

> At the school and teacher level, more attention is given to criterion referenced and diagnostic tests. If a teacher sees inconsistency between the CTBS results and results on these other instruments, she will believe the latter.
> (Evaluator)

> Individual diagnostic tools provide the basis for my judgment of program success; not the standardized tests.
> (Principal)

> The CTBS tells us by grade where the school is the weakest. We also use our curriculum tests. The results don't always match, then we go with the curriculum tests because they are more immediate and frequent. (Teacher)

> I would trust my own opinion over a test score. (Teacher)

In general, as one director put it:

> People use evaluation to support their beliefs but will not change their beliefs on disconfirming evaluation evidence.
> (Director)

Two findings connect this last point to the fact that negative standardized test results are usually ignored. First, school staff (and generally all Title I staff and parents) are happy with their programs. Second, evaluation results are looked at primarily with an eye toward confirming beliefs (see Section II). Together, then, positive results serve to reinforce existing positive feelings toward the program but negative results are ignored and, if necessary, explained away as inappropriate. When the results are negative, it does not seem to be the case that staff already knew there was a problem; in fact, the case is usually that the problem perceived is not with the program but with the tests.

The above examples illustrate the ease with which tests are written off when test results are incompatible with existing beliefs about program effectiveness. The widely publicized methodological critiques of standardized tests facilitate this process in that people who are displeased with test results can quickly call to mind "scientific" reasons for rejecting the tests. As one administrator stated:

> If the standardized test scores are negative, it's okay because everyone buys the argument that they can be discredited.
> (Administrator)

And if the problem isn't with the tests, it is with the testing conditions:

> If my judgment and the test scores tell different stories, I believe my judgment and look for explanations such as problems in giving the test or how the child felt. (Principal)

Or there is a problem with the analysis, as described below.

Important Variables Omitted

A frequent explanation for ignoring standardized test results in judging programs is that the scores are not meaningful because important background characteristics of schools or children have not been considered. Explanations of this type usually arise in the context of negative or low test results and potential comparisons with other schools or programs. For example:

> Evaluations must take into account the amount of time devoted to instruction. You can't compare programs with different amounts of instructional time or with different goals. (Teacher)

> The evaluation should have more information on the characteristics of the kids because there can be big differences between schools in socio-economic status and mobility and other things you can't measure readily. (Teacher)

> Each school in the district has different objectives. So a good school may be ranked lowest because it has harder objectives. (Parent)

> The school's drop in ranking can be explained by several factors not having to do with the program. You need to take into account the students' IQ, the number of students per staff, and the amount of instructional time per student. And some schools exclude students with low IQs when it comes to testing while others include them. (Principal)

> I would like to see more sophisticated efforts to adjust students' expected levels of achievement for a variety of factors: attendance levels, number of schools the student has attended, number of programs he has been involved in, if he uses a second language, has a learning disability or if he comes from a broken home. (Principal)

> There is great difficulty in using the same tests even if restricted to programs with the same goals because of differences in populations. For example, the bottom kids in this state are not as low as the bottom kids in New Jersey. (Evaluator)

Important Goals Not Measured

Finally, in judging their program's effectiveness, staff and parents look to information that assesses what they believe to be the most important goals of the program, usually in addition to, but occasionally instead of, achievement.

> We would like to see all kinds of alternative goals given equal place: parent involvement, student self-concept, attendance rates, library records and student enthusiasm. (Administrator)

> Test scores on the CTBS don't say very much about whether the program was successful. Test scores are less important than growth in the affective areas. (Teacher)

> Evaluation data do not show what is effective. Teacher-pupil relationships and the quality of the teacher are what makes the biggest difference. (Director)

A related concern vis-á-vis goals is that emphasis of achievement tests has narrowed the focus of Title I.

> Title I was first a poverty program; now it is entirely achievement--all activities are now instructional, as a result, in part, of using standardized tests; also because achievement tests are used as allocators at the school level. (Evaluator)

> I would like to do more than reading and math but you can't measure them so the state won't allow it. (Principal)

> We are suspicious of all hard data and see Title I shifting to reflect an obsession with testable outcomes. (Administrator)

> Why the concentration on math? Because 2 + 2 = 4. You can make clear assessments of what students know and this is much harder to do in reading. (Principal)

In summary, there are multiple reasons for the minimal use of evaluations in judging program effectiveness. Generally, the reasons that are stated reflect preferences for measures of achievement other than standardized tests, a fear of misleading comparisons, and the view that programs have multiple goals. How then do local staff and parents reach conclusions about the effectiveness of their program? This topic is discussed below.

What Information Is Used or Desired in Judging Programs?

It is impossible to isolate precisely the information on which individuals actually base their judgments of program effectiveness. Psychological theories (e.g., cognitive dissonance) suggest that there are many important variables to consider besides the availability of certain types of information. Nevertheless, since one purpose of this study is to provide a starting point for considering how evaluations might be made more useful, it is helpful to report the types of information that respondents cite when asked about program effectiveness. I consider both what information respondents claim they use and what other types of information they say they would like.

The findings reported below are grouped into three major categories: information related to cognitive growth, growth in noncognitive areas, and outcomes in areas not related to the child. The responses described under these headings were elicited primarily by asking questions such as: How would you convince me your program is a success? If you were choosing a new program, what would you consider?

Cognitive Growth

Most respondents are concerned with growth in cognitive areas, usually reading and math. Thus the tendency not to cite evaluation data as a source for program judgments is more a reflection of the perceived limitations of standardized tests than of the domain being assessed. For example,

> Standardized achievement tests provide an indicator of where children are but they do not provide very specific information about skill attainment. (Teacher)

Respondents, particularly teachers, are more likely to cite specific measures of skills as better indicators of growth than standardized achievement tests--but just as frequently they cite their own observations and experiences. Hence,

> If I were to judge a program I would first look at the written goals of the program and then at the specific goals for each child. I would want to see pre and post test scores on individual skills rather than standardized achievement tests. (Teacher)

> In judging the effectiveness of the program I look at scores on the CTBS and the Nelson and I rely on my own observations. You can just tell if a child is improving. (Teacher)

> I judge the program on the basis of my own experience. And I would back this up with the opinions of teachers and parents when the students get to the higher grades. (Principal)

Staff also prefer to base their judgments on relative rather than absolute (external) standards; that is, they want to assess progress individually as compared to where the child started:

> What matters is how far students have come--not whether they're at grade level. (Teacher)

> I would judge students' gain by where they started and amount of instruction they received. (Administrator)

Parents, understandably, rely primarily on observation of their own child. For example:

> I know whether my child can read by observing him. I have seen increases in the number of books he brings home, and the amount of time he spends reading and this is evidence to me that the program is helping my child. (Parent)

Additionally, staff frequently expressed an interest in basing judgments on the long-term impact of the program--information rarely contained in evaluation reports. For example:

> I would like to know how the students do in ninth grade as judged by their teachers. Are the gains sticking? Will they graduate? Are they interested in school? (Principal)

> I would like to see a longitudinal study over 12 years based on achievement scores. Also to know where students arrive, what their outlook is on society and on themselves. (Principal)

As some of the above quotations indicate, staff and parents are also interested in noncognitive child outcomes.

Noncognitive Outcomes

In most cases, staff and parents are interested in both cognitive and noncognitive outcomes; hence, their comments concerning program judgments cannot always be clearly sorted between the two categories. In

addition, it is generally agreed that there are few, if any, good non-cognitive measures. Usually staff and parents cite their own observations or those of others. Some examples:

> To convince someone the program was good I would use the Reading Inventory Test of Skills, even though it is not normed. Also, observation of students' motivation to see changes in personality and attitude. (Administrator)

> To see if the program was effective, I would look at four things: how well the student was doing in other classes, especially in areas which first caused them to come to the reading lab, pre and post scores on the ICRT, teacher reports, and the students' attitudes to the program. (Teacher)

> I make my judgments by looking at the children. Can the child perform? Is he at ease? Does he have a good self image? (Teacher)

> The program is successful if students get attached to their teacher, if they want to go to the program. You also know something special is going on if students not in the program want to join it. Parents get a sense of the program and communicate it to their children too. (Principal)

> The program is effective if children know what they are doing. (Teacher)

Generally, then school staff express interest in areas such as student attitude and self-concept, although formal measures are rarely cited as information sources for these areas.

Areas Not Related to the Child

In addition to judging program effectiveness on the basis of information about the participants, either cognitive or noncognitive, some staff expressed interest in the effect of the program on groups other than children. As examples:

> There are lots of ways of telling if the program is effective. Test scores are one. Others are the working relationships, the atmosphere and community attitudes toward the program--perhaps the last is most important. The community lets you know if anything is wrong. (Principal)

> Yes, the program is a success because the parents and the kids think it is helpful and the teachers are enthusiastic. (Administrator)

> After initial resistance, the staff has become supportive and you can tell it [the program] is a success when teachers say good things about the kids. (Principal)

Overall, although most staff are concerned primarily with the program's impact on children, there is interest in knowing the impact of the program on the community, parents, and staff itself. As with non-cognitive outcomes, however, little mention was made of formal ways of measuring these areas of interest.

IV USE OF EVALUATION FOR PROGRAM IMPROVEMENT*

Continuing the search for instances of evaluation use, I turn from the question of how people judge a program's effectiveness to the question of how program improvement occurs. Specifically, I investigated how decisions about program changes are made and the extent to which evaluation data are mentioned in this context. As with the discussion of uses of evaluation for judging programs, this section considers first, the limits of evaluation for program improvement and second, the types of information that are used in program improvement.

Limits of Evaluation for Program Improvement

A determination of the utility of evaluation results in program improvement must recognize that local districts have several levels of people involved in Title I, each with different information needs and decision-making authority. Administrators are concerned with the program as a whole; principals are concerned with their schools; teachers with their classes and parents with their children. Although their information needs are not necessarily mutually exclusive, they often differ substantially.

Reasons for lack of use of evaluation for program improvement can be roughly categorized in three groups: programs are quite stable, evaluations are irrelevant; and evaluations are inappropriate.

Program Stability

In considering the use of evaluation (and other information) as a basis for making decisions about program changes, I found that Title I programs are, by and large, remarkably stable.

---

*These findings are limited to the 15 districts visited by the staff. The Huron findings are not included because this topic was not pursued in their study in sufficient detail for purposes of this section.

At least 5 of the 15 districts stated this clearly. As examples:

> There really isn't much program planning going on any longer. It's more a matter of continuing to operate the way they have been going. (Director)

> Major changes in the program are never made. (Teacher)

> There have been no basic changes in Title I. The goals and methods are largely unchanged. (Director)

From the small number of examples cited when respondents were asked about program changes, it is clear that they are limited in all districts. Therefore, it should be kept in mind that the universe in which to find connections between program changes and evaluation is quite restricted.

Irrelevance of Evaluations

The finding that evaluations are considered irrelevant to program decisions is in part an inference based on staff comments concerning the overriding importance of other factors (e.g., administrative, budgetary, and political). These comments are discussed later. Other indications that evaluations are viewed as irrelevant include distrust of evaluation in general and the practical constraint of timing. For example:

> I doubt that testing provides the kind of information on which to base decisions. Title I was designed to let locals define needs. Local philosophies and priorities should shape the program. (Director)

Implying that the whole notion of evaluation is irrelevant, another director said:

> How can you evaluate when kids are starting at different places and developing at different rates? Means don't mean anything. (Director)

Finally, if the evaluation results are not available when decisions are made, they are irrelevant. In all districts there is a delay between data collection and reporting of results. Usually the evaluation is based on a spring test administration and the results are not reported until the following fall. This means, first, that program planning for the next year has already occurred--often during the spring even prior

to the administration of the posttest. Second, at the teacher level, the students who were evaluated are no longer with the same teacher. Although theoretically data one year out of date are not totally useless, some staff suggested that this timing did preclude utilization. As examples:

> Evaluation reports cannot be included in planning because plans must be submitted to the state before evaluation is available. Planning must be done at the busiest time of the year.
> (Evaluator)

> Data from the previous spring are too late to be of use, except to purchase materials. (Administrator)

Inappropriateness of Evaluations

Most staff interviewed did not speak directly to the issue of appropriateness of evaluation for making program decisions. The most obvious explanation for this is that staff do not view evaluation as a possible guide for program improvement. Instead, "evaluation" in all contexts is interpreted as a means for someone else to judge the effectiveness of the program. Thus, "evaluation" tends to be associated with accountability rather than with information for identifying strengths and weaknesses of the program. Ironically, when test scores are referred to as "needs assessment," the reaction to them can be quite different.

The only way we were able to determine any connections between evaluation and program decisions was to work backwards from program decisions previously made. We asked what changes had been made in the program and then asked why the noted changes, if any, were made. From this we were able to determine the extent to which standardized test results and other types of information played a role in the decision to make changes in the program.

Although this approach generated examples of information used in making changes (below), it did not produce spontaneous statements about why evaluations were deemed inappropriate. Therefore, my conclusion that evaluations are usually considered inappropriate for program decisions is based on inference rather than direct statements. All staff

and parents make judgments about the program, as we have noted. Because judgments provide the starting point for actions or decisions, I infer that staff would claim that evaluations are inappropriate for decisions for the same reasons that they give for their inappropriateness for program judgments--that is, because standardized tests are not convincing measures of achievement and people are concerned with outcomes not addressed by evaluations.

What Types of Information are Used in Making Decisions?

Because of the limited number of program changes, and hence the limited evidence concerning their causes, this discussion is constructed somewhat differently from the preceding ones. From the field notes for each of the 15 districts, I extracted every example of a connection between a program decision and some kind of information (defined broadly). These examples were elicited primarily in indirect fashion, through inquiries first on how the program had changed recently and then on why the changes were initiated. The examples should be interpreted in the the light of how they were collected; to wit, we took all responses at face value. We did not attempt to trace program decisions to a primary source nor to resolve conflicting explanations from different respondents in the same district. Because no program change stems from a single cause, and perceptions of causes often differ, such a task would have been impossible, For example, in one district parents were convinced that they had been responsible for the introduction of a math component; administrators, on the other hand, felt that the program had been initiated because they perceived the need and funds were available.

I identified in total approximately 35 illustrations of connections between program changes and information. The types of information cited in these illustrations can be grouped roughly into four categories: evaluation; fiscal/political; "subjective"; and "objective" with the understanding that several illustrations fall under more than one category. Examples from each category are presented below.

About one-quarter of the illustrations cited evaluation or test scores as contributing to a change in the program. Two examples are:

> From the survey information in the evaluation, I saw that some teachers in the school weren't as well informed about the Title I program as they should be so I made it a point to work with them more. (Teacher)
>
> I circle the high and low posttest scores and meet with the teachers on weaknesses to consider for next year's class. (Administrator)

Several examples in this category suggested less than a compelling connection between the test scores and the change initiated (or the change was described in such vague terms that the connection was difficult to determine). Some examples are:

> I took heed to the low scores in comprehension and did some inservice. (Principal)
>
> Test results showed that students did poorly in drawing inferences. The school responded by beefing up materials in this area. (Teacher)
>
> I look at the class results to see if anything is out of phase. I found some had dropped in math and diagnosed the problem as three different approaches being used school-wide. So I picked the one most widely used and stopped the rest. (Principal)
>
> We stopped serving three-year olds because they scored too high at the end of the year to be eligible as four-year olds. (Director)

These examples suggest that changes associated with test scores tend to be minor (excepting the last), and that the vagueness of the changes perhaps reflects the state of the art in the field of education--limited clear remedies even when a weakness has been identified.

The second category of illustrations suggests that fiscal and political considerations are at least as important as evaluation in motivating change, based on the fact that they also represent about one-fourth of all the illustrations. They tend, however, to reflect more sweeping changes. Four examples are:

> The math program came about because we had carryover funds accumulating and felt a need for a math program. (Principal)

> Most of the changes that have taken place have been shifts in the location of the program as the number of eligible students changes, as the funding increased or decreased, etc. (Director)

> Aides cost more each year so we have to eliminate some. (Principal)

> The math program was started because the state suggested it. (Director)

I suspect that budgetary and political considerations are even more influential than the total number of illustrations suggests but would tend to be mentioned less often, particularly in the context of an interview directed at local utilization of evaluation.

The third category, subjective information, includes over one-third of all the illustrations. Most of these illustrations suggest that changes were based primarily on staff observation of the program. Some examples are:

> I will expand the content area of the reading lab to science because of the success I have had using social studies materials, because science is interesting to the students, and because I hope to help them improve their work in other classes. (Teacher)

> The math program was expanded with additional personnel and diagnostic tests because we saw the need for these things; they have enhanced our basic program. (Teacher)

> We use informal evaluation (teacher experience) to modify the curriculum and use trial and error to find the right activities. The big decisions (e.g., dropping kindergarten, food, etc.) are political and administrative. If hard data are available and on the right side, things are easier to sell. (Director)

> Changes are often based on questionnaires filled out by teachers and principals and my observation. (Administrator)

Several examples in this category indicate a major concern with program manageability and teaching philosophy. As examples:

> We chose a new reading series because we felt we needed a less individualized approach and more direct contact. So we had the faculty evaluate several and also visited other schools to look at it and the scores. The faculty liked it because it gave introductions to stories and had built-in-testing. (Principal)

> We chose a new math program because the existing curriculum wasn't unified. We wanted the same program in kindergarten through sixth grade. Thus we looked at materials. Second, we looked at whether it would be effective. We did this by involving the whole staff, recommendations from the district and we knew we wanted one that didn't rely heavily on reading and that had built-in tests. (Teacher)

Finally, there were three examples that suggested use of evaluative information, but not that information reported formally in the evaluation. This is the category loosely termed "objective." The three examples are:

> One school claimed that their self-concept program was great so I measured it and found no gains. This got them to think more about what they were doing and what they expected. (Evaluator)

> We have made a major change based on three years of files from problem solving sessions with teachers. We reduced record keeping and increased small group activities. We also changed class size based on teachers' recommendations and changed materials distribution and space based on their recommendations. (Administrator)

> I got interested in unobtrusive measures to assess library use. I got a librarian to cooperate and had him checking to see if Title I kids were reading as much. They were but it tended to be the easier books. So the librarian ordered more easy books that would be of interest to the older kids--stuff that would not embarrass them. (Evaluator)

In summary, there are so few examples altogether of connections between program changes and information that it is risky to generalize from them. The fact that so few exist, especially examples in which evaluation was used, is by far the most important finding.

V DISCUSSION

At the local level, Title I evaluations are used primarily to meet requirements, to provide feedback, and to confirm (when they are positive) individual feelings toward the program. The evaluations tend not to be used either as a basis for judging the effectiveness of the program or as a guide for program decision-making. Sections II - IV consist primarily of the reasons people gave us for not using evaluation results for judgments and decision-making. To predict the impact of the proposed evaluation system, however, it is necessary to go one step beyond the stated reasons and consider underlying explanations for them, as well as constraints imposed by the context of Title I programs on the constructive use of evaluation.

Interpretation of the Findings

The stated reasons for not using evaluations tend to focus on the characteristics of the information in the evaluation. Standardized achievement test scores, the backbone of Title I evaluations, are viewed as inadequate at best for program judgments and planning. Reasons expressed for this view range from the limitations of these tests in measuring the attainment of specific skills to the omission of measures of other outcomes considered important, such as children's attitudes and parental involvement. These stated reasons imply that if the type of evaluative information reported were changed, use of the information would increase. However, a close reading of all the statements of the respondents suggests otherwise. The statements in toto suggest that there are unstated explanations for not using evaluation results as well as constraints on evaluation use imposed by the structure of the programs, both of which must be addressed directly if use of evaluation in program planning is to increase. Merely changing the type of information reported is insufficient in itself.

Section II illustrates that even some of the most common uses of evaluation are avoided when the results are negative. Uses of evaluation for public relations or for confirming one's own beliefs occur only when the results are positive. Sections III and IV demonstrate that judgments of program effectiveness and decisions about program changes rely heavily on personal, subjective information without clear expression of what is being assessed and how. On the other hand, there is evidence to suggest that standardized test scores are perceived as useful in contexts other than evaluation, such as in the selection process and for needs assessments. (These uses are referred to only briefly in the text since they were not the focal point of the study.) Together these findings suggest that there are some deeper explanations for the limited use of evaluations--reasons that go beyond the characteristics of the outcomes and measures reported--and that, therefore, the stated reasons are best understood as a reflection of the underlying reasons. Moreover, both levels of explanation must be viewed within the context of school districts and their Title I programs. I discuss first this context and then elaborate on the underlying reasons for lack of use of evaluations.

The Context of Title I Programs

Sections III and IV indicate two constraints on the use of evaluations posed by characteristics of the program and its evaluation. First, programs tend to be quite stable, thus limiting the universe in which changes are likely to be made, whether based on evaluations or not. Second, the timing of the evaluation can by itself restrict its potential utility by not meshing with the timing of program planning. Since evaluation results are generally reported after the planning has occurred, their use is at best limited to that of year-old data.

Several other constraints imposed by the structure of programs were observed, if not stated directly by respondents. Perhaps most important is the fact that, in almost every district, there is little connection between program staff and evaluation staff. This is a function of the administrative structure of the program in almost every

district. The person or persons responsible for the administration and the content of the program are not those who are responsible for the design and conduct of the evaluation. Additionally, the evaluator, particularly when he/she is external to the program staff, reports only to the Title I Director and is usually completely isolated from the program. As one evaluator stated:

> I don't know whether the test scores are useful as a basis for making changes in the program because I don't deal with the content of the program. (Evaluator)

Similarly, an external evaluator expressed distance from the program by saying:

> I am not involved with the program or process evaluation. My main audience is the Education Department of the district and the state. (Evaluator)

There were two districts in which this gap was bridged, but not without considerable effort on the part of an administrator in one and the evaluator in the other. In fact, in the latter case, the evaluator was attempting to insert the word "Planning" in the title of his office in order to communicate the relationship he is trying to establish between program and evaluation.*

Another difficulty posed by the system is that a Title I program contains multiple potential audiences for evaluation, each of which has different information needs. Title I evaluation is frequently discussed in terms of meeting Federal, state, and local needs; often overlooked in this context, however, is the fact that each LEA is a complex organization itself, with several levels from the director to curriculum supervisors or other intermediate administrators to principals and teachers, as well as parents.

---

*Even in this situation, however, the evaluator felt that a relationship between program and evaluation is impossible to establish if it is instigated only by the evaluator and not supported as well by the program's administrators.

Finally, there is a general constraint on using evaluation that stems not from the specific context of each district but from the state of the art in educational treatments. Ideally, evaluation is expected to provide evidence on the strengths and weaknesses of programs that can in turn guide planners on directions in which their programs can be improved. This ideal presupposes, however, that if a weakness is identified, there are one or more potential remedies available. The limited knowledge on what constitutes a successful strategy in educational treatments therefore limits the extent to which evaluation can be fully utilized--both from the lack of proven alternatives and from the feelings of frustration that this lack produces. This is not meant to imply that there is a magic solution just around the corner, but rather that education is a difficult if not impossible area in which to apply fully a rational model of evaluation as a guide to decision-making.

The constraints of the system are not necessarily permanent fixtures, but they do characterize the current state of affairs in the districts visited, and, I suspect, in most others. As such, they not only limit uses of evaluation directly, but also strongly affect how individuals in the system perceive evaluation. The isolation of the evaluator from the program, the relative stability of programs, and the timing of evaluations together contribute to a climate that is not conducive to viewing evaluation as a potentially constructive tool. This climate provides an important perspective for understanding why individuals in the system view evaluation as they do. This view, gleaned from looking beyond what respondents said, is described next.

Underlying Attitudes Toward Evaluation

Two facts about the state of mind of local staff suggest strongly that, regardless of the type or quality of the evaluation data, the data are not likely to be favorably received and hence not used. The first is the narrow and usually negative way in which evaluation is perceived and the second is the strong motivation of individuals to protect their basic beliefs.

Put in the simplest terms by one evaluator: "Evaluation is a dirty word." In general, evaluation is viewed as a set of procedures designed to provide one's superiors with information on which to judge the program's success, on the basis of criteria defined by the superiors. Evaluation is therefore more likely to be associated with the threat of accountability to someone else than with its potential as a useful source of information for one's self. To the extent that the evaluation questions and criteria for success are imposed externally and that the evaluation is conducted primarily to meet externally imposed requirements, this negative view of evaluation is reinforced by actual experience. Furthermore, its threatening nature is exacerbated by the psychological distance between evaluation and program staff. As long as evaluation is viewed in this narrow and essentially threatening way, it is doubtful that the information it contains will be used, regardless of its characteristics.

The second state of mind can be characterized as the "true-believer" syndrome. It is common knowledge that an individual deeply committed to a particular belief is not likely to change that belief merely because "objective" evidence against the belief is presented. Politics and religion abound with relevant examples. This is not to imply that local Title I staff and parents are akin to religious zealots, but they are by and large strongly committed to their programs. When people invest their time and energy in a cause they view as worthy, they will seek out and readily accept evidence that their work has not been in vain. Likewise, they will ignore or explain away information that suggests they have failed. Title I staff, particularly those involved daily in implementing the program, often invest considerable energy in their work because they view it as important and worthwhile. Therefore, it is not surprising that they interpret evaluation results selectively, accepting the positive and rejecting the negative. As one director said, "We are successful even if we can't show it on paper."

Conclusion

From this analysis, I conclude that changing the type or quality of information contained in the evaluations will not, by itself, affect

the level of evaluation utilization. But simply changing the nature of the information is the focal point of the USOE evaluation models and the primary role of the TACs, which is to assist in the implementation of the models. The models address only the "symptoms," that is, technical weaknesses of the outcome measures and procedures for data collection and analysis. I suggest that this approach--and any approach that focuses exclusively on the information contained in the evaluations--cannot by itself significantly affect local use of evaluation. Instead, changes in the evaluation system designed to increase local utilization must address the underlying reasons for lack of use, including individual attitudes and beliefs about the program and evaluation. At the same time, such a system must address those elements of the context amenable to change that reinforce existing negative views toward evaluation.

Tackling the area of attitude change is obviously far more challenging than merely changing the test or metric, but it is not beyond reach. The fact that there are even a few instances of evaluation use in program decisions suggests that increased use of evaluation is possible. This fact, together with an understanding of the impediments to use of data, point to some promising directions for the shaping of a Federal strategy that can increase local use of evaluation.

Implications for Policy

Of the two current Federal strategies for Title I evaluation, an independent national study and the three-tiered reporting scheme, only the latter has the potential to encourage local use of evaluation. Since independent national studies are generally agreed to be the best source for providing evidence of the national impact of Title I, it should be possible to emphasize local use of evaluation in the three-tiered reporting system without sacrificing a source for national impact data. Therefore, the implications discussed below take the form of recommendations for radically changing the emphasis of the three-tiered reporting system to one that encourages local use of the evaluation data.

First of all, any strategy designed to increase local use of evaluation must be grounded in a Federal commitment to this goal--a commitment which must be understood and shared by the states and communicated clearly to local districts. As long as districts collect data primarily or exclusively for state and Federal use, they are unlikely to change their views toward evaluation. This suggests, at the least, that deadlines for evaluation reporting should be coordinated with the local planning cycle.

Second, districts need assistance in increasing communication and cooperation between program staff and evaluation staff. Our visits suggest that the provision of feedback can be used as one way to facilitate understanding between program and evaluation staff. However, the information fed back must be designed in a way that makes it clearly understandable to staff and parents and must address the different needs of different levels within a district. For example, a curriculum supervisor overseeing a program in six schools views the program from a different perspective and has information needs different from those of a classroom teacher. Additionally, the findings suggest that results should be presented in person if they are to be clearly understood and hence utilized by staff.

Finally, Title I staff and parents need assistance in developing an understanding of the constructive role that evaluation can play and in acquiring certain types of nontechnical evaluation skills. Incorporating evaluation information into planning and decision making is not an automatic process, yet it is one in which local staff have received little if any training. In particular, they need assistance in learning how to ask their own evaluation questions. If the primary purpose of evaluation remains that of answering questions imposed externally, the evaluation will continue to be potentially more threatening than helpful. If, on the other hand, the evaluation responds to questions about program effectiveness that the staff have expressed interest in, the potential for using the results should increase dramatically. Until Title I staff and parents come to see Title I as a program to be improved continually based in part on evaluation, the evaluation results, even if technically sound, will fall on deaf ears.

The areas described above are limited to the ones that I feel are amenable to change through Federal policy and the provision of technical assistance. However, they do provide a starting point for designing an evaluation strategy whose primary aim is to provide local staff and parents with information of use to them in improving their programs. Additionally, the more of these issues that are addressed concurrently by an evaluation strategy, the greater the likelihood of success. Treating each cause of lack of utilization separately is less likely to affect basic attitudes toward evaluation than treating as many as possible concurrently.

Until local staff view evaluation in a positive light, effort devoted exclusively to the development of technically sound data will be wasted. In the absence of use of evaluation information, it is impossible to determine the extent to which the types of measures employed facilitate or impede use of the results. This is not to imply that the issue of measures should be ignored. Use of information is determined jointly by the characteristics of the information and the characteristics of potential users. Furthermore, the characteristics of the information can, in theory, affect the attitudes of the audience. From the comments of respondents, I suspect that measures other than standardized achievement tests should be included in the evaluation, at the least. Given the current state of affairs, however, the issue of outcome measures is far less important than that of redesigning the evaluation strategy to encourage local use through addressing the impediments discussed above.

The current technical assistance strategy, if redirected, can serve as a powerful force in changing how evaluation is perceived and thereby increase evaluation use at the local level. To accomplish this goal, however, technical assistance must be redesigned to communicate a new view of the role of evaluation and to develop skills such as generating one's own evaluation questions. As long as technical assistance is defined narrowly as a way of telling local staff "how to improve the quality of their data," it will not affect local use of evaluation.

As a final note, I would like to add a context for these findings that extends beyond the education community. The failure to use information as rational models would predict is the rule rather than the exception among decisionmakers in every walk of life that has been investigated. Thus, the portrait of educators as irrational that might be drawn from this report could equally well describe their counterparts among lawyers, physicians, or policymakers in general. The point is not to decry irrationality, but rather to direct resources toward activities that have the potential to increase the rational component of decisions.

REFERENCES

Cohen, D.K. and M. Garet, "Reforming Educational Policy with Applied Social Research," Harvard Educational Review, Volume 45, No. 1, pp. 17-43 (1975).

Frankel, C., Controversies and Decisions (Russell Sage Foundation, New York, 1976).

GAO, "Problems and Needed Improvements in Evaluating Office of Education Programs," Report to the Congress, General Accounting Office, Comptroller General of the United States (Septemeber 1977).

McLaughlin, M.W., Evaluation and Reform: The Elementary and Secondary Educacion Act of 1975, Title I (Ballinger, Cambridge, Massachusetts, 1975).

RMC, "Further Documentation of State ESEA Title I Reporting Models and Their Technical Assistance Requirements, Phase I," Report to the Office of Education, RMC Research Corporation, Mountain View, California (1976).

USCAN, "Legislative History of P.L. 93-380: Education Amendments of 1974," U.S. Congressional and Administrative News (1974).

Weiss, C., Using Social Research in Public Policy Making (D.C. Heath, Lexington, Massachusetts, 1977).